不完备混合决策系统的三支决策模型与规则获取方法 *

钱文彬 [a, b]，彭莉莎 [a†]，王映龙 [a]，段德林 [b]

(江西农业大学 a. 计算机与信息工程学院; b. 软件学院, 南昌 330045)

摘　要：现有三支决策主要针对各类完备信息系统或不完备单一型信息系统进行研究，而现实应用领域中数据往往呈现不完备性和复杂性等特征，为此，构建面向不完备混合决策系统的三支决策模型与规则获取方法。首先，计算不完备混合数据的完备邻域容差类，并将其代替等价类计算三支决策模型的条件概率；然后，根据扩展的损失函数区间概念获取各对象在"乐观""折中"和"悲观"决策下的不同阈值，进而针对不完备混合决策系统构造三种决策风险下的三支决策模型；最后，通过理论分析和医疗诊断实例详细分析了算法的有效性和可解释性，并通过实验比较和分析可知：所构模型较其他已有模型的分类过程更加合理有效，同时该模型也扩充了三支决策模型和知识发现的理论与应用研究。

关键词：粗糙集；三支决策；不完备混合数据；规则获取；粒计算

中图分类号：TP18　　doi: 10.19734/j.issn.1001-3695.2018.10.0806

Three-way decisions model and rule acquisition method for incomplete composite decision system

Qian Wenbin[a, b], Peng Lisha[a†], Wang Yinglong[a], Duan Delin[b]

(*a. School of Computer & Information Engineering, b. School of Software, Jiangxi Agricultural University, Nanchang* 330045, *China*)

Abstract: The existing three-way decisions mainly focus on various types of complete information systems or incomplete single-type information systems, while the data in the real application fields often exhibit incompleteness and complexity. Therefor this paper proposed the three-way decisions model and rule acquisition method for incomplete composite decision system based on a new complete neighborhood tolerance relation. First, the algorithm designed the new complete neighborhood tolerance classes under the incomplete mixed data, which can substitute for equivalence classes to calculate the conditional probabilities in the classical three-way decisions model. Then, the algorithm also improved the concept of the loss function interval to obtain different thresholds of each object under the "optimistic", "compromising" and "pessimistic" decisions and constructed there three-way decisions models and rule acquisition method under three decision-making risks for the incomplete composite decision system. Finally, a series of theoretical analysis and a medical diagnosis example analyzed the validity and interpretable of the algorithm in detail, and the experimental results based on different datasets show that the proposed model outperformed the state-of-the-art models in terms of classification process and expand the theory and application of three-way decisions and knowledge discovery.

Key words: rough set; three-way decisions; incomplete composite data; rule acquisition; granular computing

0　引言

Yao 教授提出的三支决策[1,2]提供了对粗糙集理论的深入理解和其在粒计算中的实际应用，现已成为知识发现领域中一个重要的研究方向。三支决策的延迟策略可一定程度提高分类准确率和降低误分类损失。近几年，三支决策扩展了对知识系统的研究范围，并在数据分析和处理中发挥着关键作用，在理论扩充和实际应用也获得极大发展。理论扩充主要有三支聚类[3]、三支概念分析[4]、三支属性约简[5,6]、序贯三支决策[7]、三支+X（X：模糊集、区间集等）[8~10]等。应用领域主要有推荐系统[11]、人脸识别[12]、恶意软件分析[13]和垃圾邮件过滤[14]等。

在实际应用中，由于数据采集限制或测量误差等原因，使得在知识获取时往往存在数据不完备的现象。经典粗糙集无法直接处理不完备数据，现有针对不完备符号型数据的处理方法有基于容差关系[15]、非对称相似关系[16]、限制容差关系[17]、完备容差关系[18]、改进的完备容差关系[19]等多种扩充粗糙集模型，而针对不完备连续型信息系统的处理方法则大多基于邻域容差关系模型。在对不完备符号型信息系统的三支决策研究中，Liu 等人[20]利用概率分布计算近似类，并利用区间概念计算各对象的三支阈值，为该类不完备符号型信息系统构建了一种新的三支决策模型。在现实应用领域中，一个信息系统中同时存在多种类型的数据是一种普遍现象。例如，在高校信息系统中，学生所选课程和任课老师等是符号型数据，学号和成绩等则是连续型数据；在医疗检测结果中，性别和血型等是符号型数据，血糖和血压等是连续型数据，包含伴有不同程度缺失的此类数据的系统都是不完备混合型信息系统。现有针对不完备混合信息系统的处理有拆分

收稿日期：2018-10-15；**修回日期：**2018-12-07　**基金项目：**国家自然科学基金资助项目（61502213，61662023，71461013）；江西省自然科学基金资助项目（20161BAB212049）

作者简介：钱文彬（1984-），男，副教授，硕导，博士，主要研究方向为粒计算、三支决策、知识发现；彭莉莎（1994-），女（通信作者），硕士研究生，主要研究方向为三支决策、数据挖掘（peng_lisha@163.com）；王映龙（1970-），男，教授，硕导，博士，主要研究方向为计算智能、知识发现；段德林（1997-），本科，主要研究方向为算法设计、数据挖掘.

处理法和通用处理法，姚晟等人[21]通过考虑属性值的概率分布分别为不完备的连续型数据和离散型数据构造不确定性度量方法；Zhao 等人[22]提出邻域容差关系用于计算不完备混合数据的近似类，并构建邻域条件熵作为评估标准进行属性约简；然而在三支决策理论模型中，对于同时包含连续型、整型和符号型数据的不完备混合决策系统的研究相对尚少。

为此，本文为不完备混合决策系统构建三支决策模型和规则获取方法。对于条件概率的计算，以改进的完备容差关系为基础，定义一种新的完备邻域容差关系并将其用于计算不完备混合数据的近似类，进而有效替代等价类计算条件概率；针对阈值的获取，由于未知属性值的不确定性和不稳定性，每个对象都应设置不同的阈值，为此，改进文献[20]中近似类的损失函数区间值获取方法，首先为各对象设置损失函数区间，然后通过改进"乐观""折中"和"悲观"损失函数具体值设置方式计算各对象的"乐观""折中"和"悲观"三支阈值。本文模型的主要优势是能利用通用的数学范式计算同时包含不完备符号型、整型和连续型数据的对象的近似类，且可根据三种风险偏好分别为各不完备对象设置不同的阈值，进而制定三种风险偏好下的三支决策规则。通过理论分析和医疗诊断的实例证明了本文所构模型的有效性和可行性，且通过 UCI 数据集的实验分析了参数的变化对该模型的影响，对比邻域容差关系等其他三种方法，本文方法的分类过程更合理，具有更好的可解释性和扩展性。

1 基础知识

给定不完备混合决策系统 $IDS = U,C,D$，其中：U 为对象集；C 为条件属性集；D 为决策属性。令 $\forall B \subseteq C$，且 $B = B_C \cup B_N$，B_C 为符号型属性子集，B_N 为连续型属性子集。预先设定"*"代表未知属性值（空值），通常未知属性值分为遗漏值和缺失值，对于遗漏值而言，未知值只是遗漏但确实存在，且可与任意同类属性值进行比较，即各对象具有潜在的完备信息；而对于缺失值而言，未知值可能是不可获取的或者根本不存在的数据，不能与其他属性值相比较。本文只对遗漏值进行研究。

为处理不完备混合决策系统中的连续型数据，将邻域概念用于容差关系、限制容差关系、完备容差关系和改进的完备容差关系下的近似求解。

定义 1[5] 给定不完备混合决策系统 $IDS = U,C,D$，设混合属性子集 $B \subseteq C$，$\forall a_k \in B$，$a_k(x)$ 和 $a_k(y)$ 分别表示对象 x 和 y 在属性 a_k 上的属性值，则对 $\forall x,y \in U$，邻域距离函数定义为

$$\Delta_B(x,y) = \left(\sum_{k=1}^{n} |a_k(x) - a_k(y)|^p \right)^{1/p}$$

其中：当 $p=1$ 时，$\Delta_B(x,y)$ 为曼哈顿距离；当 $p=2$ 时，$\Delta_B(x,y)$ 为欧氏距离。

1.1 邻域容差关系

在容差关系[15]下，"*"与任意值相等。结合邻域概念定义邻域容差关系：

定义 2 给定不完备混合决策系统 $IDS = U,C,D$，令 $\forall B \subseteq C$，$B = B_C \cup B_N$，δ 为邻域参数，则对 $\forall x,y \in U$ 在 B 下的邻域容差关系定义为

$$NT_B(x,y) = \left\{ (x,y) \in U^2 \mid \forall_{a_k \in B} \left(\begin{array}{c} a_k(x) = * \vee a_k(y) = * \\ \vee \left(\begin{array}{c} (a_k \in B_C \to \Delta_{a_k}(x,y) = 0) \wedge \\ (a_k \in B_N \to \Delta_{a_k}(x,y) \le \delta) \end{array} \right) \end{array} \right) \right\}$$

NT_B 满足自反性和对称性，但不满足传递性。对 $\forall x \in U$ 在混合属性子集 B 下的邻域容差类定义为

$$NT_B(x) = \{ y \in U | y \in NT_B(x,y) \}$$

定义 3 给定不完备混合决策系统 $IDS = U,C,D$，令 $\forall X \in U$，则在混合属性子集 B 下，基于邻域容差关系的三支决策规则定义为

若 $\alpha \le P(X|NT_B(x)) \le 1$，则 $x \in POS_B^{NT}(X)$

若 $\beta < P(X|NT_B(x)) < \alpha$，则 $x \in BND_B^{NT}(X)$

若 $0 \le P(X|NT_B(x)) \le \beta$，则 $x \in NEG_B^{NT}(X)$

其中：$P(X|NT_B(x)) = \frac{|X \cap NT_B(x)|}{|NT_B(x)|}$，为 NT_B 属于 X 的条件概率；α 和 β 为三支决策阈值。

仔细分析发现，在邻域容差关系下，认为"*"与任意属性值都相似或相等，虽然这种方式几乎不会影响对缺失数据较少的信息系统的划分，但会将那些拥有较多未知属性值的对象也划分到一个邻域容差类中，且对象的未知属性值越多，被划分到同一个类中的概率越大。例如，对象 $x = (0,*,0.1,*,0.4,*,0.7,*,1)$ 与对象 $y = (*,1,*,0.8,*,0.5,*,0.2,*)$，因为这两个对象实际上相似的概率很小，所以这种划分方式过于宽松，划分粒度过粗。

1.2 限制邻域容差关系

限制容差关系[17]稍微弥补了容差关系[15]和相似关系[16]的不足。下面结合邻域概念定义限制邻域容差关系：

定义 4 给定不完备混合决策系统 $IDS = U,C,D$，令 $\forall B \subseteq C$，$B = B_C \cup B_N$，δ 为邻域参数，则对 $\forall x,y \in U$ 在 B 下的限制邻域容差关系定义为

$$NL_B(x,y) = \left\{ (x,y) \in U^2 | \forall_{a_k \in B} \left(\begin{array}{c} a_k(x) = * \vee a_k(y) = * \\ (P_B(x) \cap P_B(y) \ne \varnothing) \wedge \\ \vee \left(\begin{array}{c} (a_k \in B_C \to \Delta_{a_k}(x,y) = 0) \wedge \\ (a_k \in B_N \to \Delta_{a_k}(x,y) \le \delta) \end{array} \right) \end{array} \right) \right\}$$

其中：$P_B(x) = \{ a_k | a_k \in B \wedge a_k(x) \ne * \}$，$NL_B$ 满足自反性和对称性，但不满足传递性。对 $\forall x \in U$ 在混合属性子集 B 下的限制邻域容差类定义为

$$NL_B(x) = \{ y \in U | y \in NL_B(x,y) \}$$

定义 5 给定不完备混合决策系统 $IDS = U,C,D$，令 $\forall X \in U$，则在混合属性子集 B 下，X 基于限制邻域容差关系的三支决策规则定义为

若 $\alpha \le P(X|NL_B(x)) \le 1$，则 $x \in POS_B^{NL}(X)$

若 $\beta < P(X|NL_B(x)) < \alpha$，则 $x \in BND_B^{NL}(X)$

若 $0 \le P(X|NL_B(x)) \le \beta$，则 $x \in NEG_B^{NL}(X)$

其中：$P(X|NL_B(x)) = \frac{|X \cap NL_B(x)|}{|NL_B(x)|}$ 为 $NL_B(x)$ 属于 X 的条件概率；α 和 β 为三支决策阈值。

两个对象在满足邻域容差关系的基础上还必须满足至少一个已知属性值相等或相似，才符合限制邻域容差关系，然而这种划分方式对于缺失程度大的对象仍然存在较大的缺陷。例如，对象 $x = (0.5,*,*,*,*,*,*,*,1)$ 与对象 $y = (0.51,*,*,*,*,*,*,*,1)$，$x$ 与 y 存在 80%的未知值，实际相似的概率很小，却被划分到同一个限制邻域容差类，显然这种划分方式也过于宽容，会导致划分粒度过大。

1.3 完备邻域容差关系

由于不同的信息系统的完备度各不相同，所以完备容差关系[18]在限制容差关系的基础上考虑了单个信息系统内对象之间和多个信息系统之间的差异性，即在一个信息系统中的两个对象，基于完备容差关系被认为是可分辨的，但在另一个信息系统中可能被认为是不可分辨的。

定义 6 给定不完备混合决策系统 $IDS = U,C,D$，则对 $\forall x_i \in U$，其完备度定义为

$$\gamma(x_i) = |P(x_i)| / |C|$$

其中：$P(x_i)=\{a|a\in C\wedge a(x_i)\neq *\}$ 表示非空属性集；$|*|$ 表示集合“*”的基数，则整个决策系统的完备度定义为

$$\gamma=\left(\sum_{i=1}^{|U|}\gamma(x_i)\right)/|U|$$

由以上范式可知，$\gamma(x_i)\in[0,1]$，$\gamma\in[0,1]$。

定义 7 给定不完备混合决策系统 $IDS=U,C,D$，令 $\forall B\subseteq C$，$B=B_C\cup B_N$，δ 为邻域参数，则对 $\forall x,y\in U$ 在 B 下的完备邻域容差关系定义为

$$NM_B^{\delta}(x,y)=\left\{(x,y)\in U^2|\forall_{a_k\in B}\left(\begin{array}{c}(P_B(x)\neq\varnothing)\wedge(P_B(y)\neq\varnothing)\\ \wedge\left(\begin{array}{c}\frac{|P_B(x)\cap P_B(y)|}{\min(|P_B(x)|,|P_B(y)|)}\geq\gamma\wedge\\(a_k\in B_C\to\Delta_{a_k}(x,y)=0)\wedge\\(a_k\in B_N\to\Delta_{a_k}(x,y)\leq\delta)\end{array}\right)\end{array}\right)\right\}$$

在完备容差关系下，约定 $x=(*,*,*,*,*,*)$ 不与任何对象相似。NM_B^{δ} 具有自反性和对称性，但不具备传递性。对 $\forall x\in U$ 在混合属性子集 B 下的完备邻域容差类定义为

$$NM_B^{\delta}(x)=\{y\in U|y\in NM_B^{\delta}(x,y)\}$$

定义 8 给定不完备混合决策系统 $IDS=U,C,D$，令 $\forall X\in U$，则在混合属性子集 B 下，X 基于完备邻域容差关系的三支决策规则定义为

若 $\alpha\leq P(X|NM_B(x))\leq 1$，则 $x\in POS_B^{NM}(X)$

若 $\beta<P(X|NM_B(x))<\alpha$，则 $x\in BND_B^{NM}(X)$

若 $0\leq P(X|NM_B(x))\leq\beta$，则 $x\in NEG_B^{NM}(X)$

其中：$P(X|NM_B(x))=\frac{|X\cap NM_B(x)|}{|NM_B(x)|}$ 为 $NM_B(x)$ 属于 X 的条件概率；α 和 β 为三支决策阈值。

假设 $\pi=\frac{|P_B(x)\cap P_B(y)|}{min(|P_B(x)|,|P_B(y)|)}$，经分析发现，当两个对象的已知属性存在包含关系时，π 的值总会为 1，它必然大于或等于整个信息系统的完备度 γ。例如，对象 $x=(*,*,*,*,0.5,*,*,*,1,*)$ 与对象 $y=(2,*,*,*,0.51,*,3,*,1,*)$，因为 $\pi(x,y)=1$，其大于等于 γ 的概率为 100%，在完备邻域容差关系下，x 与 y 被认为是不可分辨的。但实际上，它们相似的可能性很小。因此，完备邻域容差关系对缺失数据的处理也存在一定不合理性，且当信息系统的不完备度较大时，其不合理性更明显。

2　不完备混合决策系统的三支决策模型

不完备混合决策系统的三支决策模型的核心步骤为：定义新的完备邻域容差关系用于计算不完备混合数据的近似类，进而计算条件概率，然后通过新定义的损失函数区间概念计算“乐观”“折中”和“悲观”三支阈值。

2.1　一种新的完备邻域容差关系近似度量

定义 9 给定不完备混合决策系统 $IDS=U,C,D$，令 $\forall B\subseteq C$，$B=B_C\cup B_N$，B_C 为符号型属性子集，B_N 为连续型属性子集，则对 $\forall x,y\in U$，其新的完备邻域容差关系定义为

$$NR_B(x,y)=\left\{(x,y)\in U^2|\begin{array}{c}\forall_{a_k\in B}(a_k(x)=*\vee a_k(y)=*)\\ \wedge min\left(\frac{|P_B(x)\cap P_B(y)|}{min(|P_B(x)|,|P_B(y)|)},\frac{\gamma(x)+\gamma(y)}{2}\right)\geq\gamma\end{array}\right\}$$

其中：

$$P_B(x)\cap P_B(y)=\begin{cases}\{a_k|a_k(x)\doteq a_k(y)\},a_k\in B_C\\ \{a_k|\Delta_{a_k}(x,y)\leq\delta_{a_k}\},a_k\in B_N\end{cases}$$

$$P(x)=\{a_k|a_k\in B\wedge a_k(x)\neq *\}$$

$$\Delta_{a_k}(x,y)=|a_k(x)-a_k(y)|。$$

δ_{a_k} 为属性 a_k 上的邻域参数，为保证新的完备容差邻域类的计算的客观性，δ_{a_k} 由 a_k 的标准差 std_{a_k} 和参数 ϑ 表示，即 $\delta_{a_k}=\frac{std_{a_k}}{\vartheta}$；此处约定 $\frac{0}{0}=0$，即 $x=(*,*,\ldots,*)$ 不与任何对象相似。

$NR_B(x,y)$ 满足自反性和对称性，不满足传递性，对 $\forall x\in U$，其在不完备混合属性子集 B 下的新完备邻域容差类定义为

$$NR_B(x)=\{y|y\in NR_B(x,y)\}$$

同样，对于定义 7 中给出的对象 $x=(*,*,*,*,0.5,*,*,*,1,*)$ 与对象 $y=(2,*,*,*,0.51*,3,*,1,*)$，在新的完备邻域容差下，计算得到 $\frac{\gamma(x)+\gamma(y)}{2}=0.3$，在一个包含了 x 和 y 的系统中，γ 值有 70% 的概率在 0.3~1 之间，而这意味着 $\frac{\gamma(x)+\gamma(y)}{2}$ 小于 γ 的概率为 70%，即 x 与 y 相似的概率仅为 30%，所以 x 与 y 被认为是不相似的。因此，对于不完备混合数据的处理，新的完备邻域容差关系更加客观合理。

定理 1 给定不完备混合决策系统 $IDS=U,C,D$，若新完备邻域容差类由 $P_B(x)\cap P_B(y)$ 决定，则当邻域参数 $\delta_{a_k}^{\delta_i}\leq\delta_{a_k}^{\delta_j}$ 时，有 $NR_B^{\delta_i}(x,y)\subseteq NR_B^{\delta_j}(x,y)$。

证明 根据定义 9，对于 $\forall y\in NR_B^{\delta_i}(x,y)$，都有 $\Delta_{a_k}(x,y)\leq\delta_{a_k}^{\delta_i}$，又因为 $\delta_{a_k}^{\delta_i}\leq\delta_{a_k}^{\delta_j}$，所以 $\Delta_{a_k}(x,y)\leq\delta_{a_k}^{\delta_j}$，即 $y\in NR_B^{\delta_j}(x,y)$，所以可以得到 $NR_B^{\delta_i}(x,y)\subseteq NR_B^{\delta_j}(x,y)$。

定理 2 给定不完备混合决策系统 $IDS=U,C,D$，$NT_B(x)$ 为邻域容差类，$NL_B(x)$ 为限制邻域容差类，$NM_B(x)$ 完备邻域容差类，$NR_B(x)$ 为新完备邻域容差类，若 $\forall_{x\in U}(P(x)\neq\varnothing)$，都有

$$NR_B(x)\subseteq NM_B(x)\subseteq NL_B(x)\subseteq NT_B(x)$$

证明 由定义 2、4、7 和 9 可证。

定义 10 给定不完备混合决策系统 $IDS=U,C,D$，在新的完备容差邻域关系下，对 $\forall X\in U$，其在不完备混合属性子集 B 下的上下近似定义为

$$X_B^{NR}=\{x|x\in U\wedge NR_B(x)\subseteq X\}$$

$$X_{NR}^{B}=\{x|x\in U\wedge NR_B(x)\cap X\neq\varnothing\}=\bigcup_{x\in X}NR_B(x)$$

定理 3 给定不完备混合决策系统 $IDS=U,C,D$，NT 为邻域容差关系，NL 为限制邻域容差关系，NM 完备邻域容差关系，NR 为新的完备邻域容差关系，若 $\forall_{x\in U}(P(x)\neq\varnothing)$，则对 $\forall X\in U$ 都有

$$X_B^{NT}\subseteq X_B^{NL}\subseteq X_B^{NM}\subseteq X_B^{NR}$$

$$X_{NR}^{B}\subseteq X_{NM}^{B}\subseteq X_{NL}^{B}\subseteq X_{NT}^{B}$$

证明 a)：若存在 $x\in X_B^{NT}$，则有 $NT_B(x)\subseteq X$，由定理 1 可知，$NL_B(x)\subseteq NT_B(x)\subseteq X$，即可得到 $x\in X_B^{NL}$，因此 $X_B^{NT}\subseteq X_B^{NL}$，同理可证 $X_B^{NL}\subseteq X_B^{NM}$ 和 $X_B^{NM}\subseteq X_B^{NR}$，所以 $X_B^{NT}\subseteq X_B^{NL}\subseteq X_B^{NM}\subseteq X_B^{NR}$。

证明 b)：由定理 1 可知 $NR_B(x)\subseteq NM_B(x)\subseteq NL_B(x)\subseteq NT_B(x)$，又因为 $X_{NR}^{B}=\cup NR_B(x)$，所以 $X_{NR}^{B}\subseteq X_{NM}^{B}\subseteq X_{NL}^{B}\subseteq X_{NT}^{B}$。

2.2　新的完备邻域容差关系三支决策模型

在决策粗糙集中，条件概率由等价类 $[x]$ 计算而来，且 $[x]$ 中的对象共用相同的损失函数 $\lambda_{\cdot\cdot}(\cdot=P,B,N)$。在本文中，条件概率由新完备邻域容差类 $NR_B(x)$ 计算而来，而 $NR_B(x)$ 中的对象都包含未知值，所以每个对象都应设置不同的损失函数。在不能准确地估计损失函数的情况下，可通过定义损失函数

区间概念，利用区间值 $\lambda_{\cdot\cdot}=[\lambda_{\cdot\cdot}^{-},\lambda_{\cdot\cdot}^{+}]$[20]计算损失函数进而计算阈值，因此构建损失函数区间表如表 1 所示。

表 1　损失函数区间

Table 1　Loss function interval

决策	$X(P)$	$\neg X(N)$
接受：a_P	$\lambda_{PP}=[\lambda_{PP}^{-},\lambda_{PP}^{+}]$	$\lambda_{PN}=[\lambda_{PN}^{-},\lambda_{PN}^{+}]$
延迟：a_B	$\lambda_{BP}=[\lambda_{BP}^{-},\lambda_{BP}^{+}]$	$\lambda_{BN}=[\lambda_{BN}^{-},\lambda_{BN}^{+}]$
拒绝：a_N	$\lambda_{NP}=[\lambda_{NP}^{-},\lambda_{NP}^{+}]$	$\lambda_{NN}=[\lambda_{NN}^{-},\lambda_{NN}^{+}]$

表中：$\forall X\in U$、$\lambda_{\cdot\cdot}^{-}\le\lambda_{\cdot\cdot}^{+}$、$\{\lambda_{PP},\lambda_{BP},\lambda_{NP}\}$ 分别表示当对象 x 属于对象子集 X 时，接受、延迟和拒绝 x 于 X 类所造成的损失；$\{\lambda_{PN},\lambda_{BN},\lambda_{NN}\}$ 分别表示当 x 不属于 X 时，接受、延迟、拒绝 x 于 X 类所造成的损失，故作合理假设：$\lambda_{PP}\le\lambda_{BP}\le\lambda_{NP}$，$\lambda_{NN}\le\lambda_{BN}\le\lambda_{PN}$。

预先设定损失函数参数 μ，将损失函数区间转换为具体的损失函数值：$\lambda_{\cdot\cdot}=(1-\mu)\lambda_{\cdot\cdot}^{-}+\mu\lambda_{\cdot\cdot}^{+}$。其中，当 $\mu=0$ 时，$\lambda_{\cdot\cdot}$ 偏小，代表偏好风险设置法，即“乐观”设置法；当 $\mu=1$ 时，$\lambda_{\cdot\cdot}$ 偏大，代表厌恶风险设置法，即“悲观”设置法；当 $\mu=0.5$ 时，$\lambda_{\cdot\cdot}$ 为区间平均值，代表风险折中设置法，即“折中”设置法。由于文献[20]中近似类的损失函数区间获取时可能会出现没有交集或者没有并集的情况，所以本文定义新的损失函数区间概念如下：

定义 11 给定伴有损失函数区间的不完备混合决策系统 $IDS=U,C,D,\lambda_{\cdot\cdot}$，则对 $\forall x\in U$，其在“乐观”“折中”决策和“悲观”决策下的损失函数 $\lambda_{\cdot\cdot}(\cdot=P,B,N)$ 定义为

“乐观”损失函数值：$\lambda_{\cdot\cdot}^{\curlyvee}(x)=\dfrac{\sum_{x_i\in NR_B(x)}\lambda_{\cdot\cdot}^{-}(x_i)}{|NR_B(x)|}$；

“折中”损失函数值：$\lambda_{\cdot\cdot}^{\succ}(x)=\dfrac{\sum_{x_i\in NR_B(x)}\left(\dfrac{\lambda_{\cdot\cdot}^{-}(x_i)+\lambda_{\cdot\cdot}^{+}(x_i)}{2}\right)}{|NR_B(x)|}$；

“悲观”损失函数值：$\lambda_{\cdot\cdot}^{\curlywedge}(x)=\dfrac{\sum_{x_i\in NR_B(x)}\lambda_{\cdot\cdot}^{+}(x_i)}{|NR_B(x)|}$；

以上范式中的上标符号分别表示：“$\curlyvee$”—“乐观”，“$\succ$”—折中，“$\curlywedge$”—“悲观”。为使文章简洁，在下文中所出现的这些符号均表示该含义。

三支决策的“三分—三决策”模型的基本框架是通过一对阈值将论域划分为三个互不相交的域，即 POS—正域、BND—边界域和 NEG—负域，针对这三个域中的对象分别采取接受、不承诺和拒绝的策略。为此，基于新的完备邻域容差关系为不完备混合决策系统构建三支决策模型。

定义 12 给定不完备混合决策系统 $IDS=U,C,D,\lambda_{\cdot\cdot}$，令 $\alpha_{\curlyvee}$ 和 $\beta_{\curlyvee}$ 为基于损失函数 $\lambda_{\cdot\cdot}(\cdot=P,B,N)$ 计算而来的乐观三支阈值，当 $\alpha_{\curlyvee}\ge\beta_{\curlyvee}$ 时，在新的完备容差邻域关系下，对 $\forall X\in U$，其在不完备混合属性子集 B 下，依据贝叶斯风险最小化决策准则制定的“乐观”三支决策模型为

若 $\alpha_{\curlyvee}\le P(X|NR_B(x))\le 1$，则 $x\in POS_B^{NR}(X)$

若 $\beta_{\curlyvee}<P(X|NR_B(x))<\alpha_{\curlyvee}$，则 $x\in BND_B^{NR}(X)$

若 $0\le P(X|NR_B(x))\le\beta_{\curlyvee}$，则 $x\in NEG_B^{NR}(X)$

其中：$P(X|NR_B(x))=\dfrac{|X\cap NR_B(x)|}{|NR_B(x)|}$。

$$\alpha_{\curlyvee}=\frac{(\lambda_{PN}^{\curlyvee}-\lambda_{BN}^{\curlyvee})}{(\lambda_{PN}^{\curlyvee}-\lambda_{BN}^{\curlyvee})+(\lambda_{BP}^{\curlyvee}-\lambda_{PP}^{\curlyvee})},$$

$$\beta_{\curlyvee}=\frac{(\lambda_{BN}^{\curlyvee}-\lambda_{NN}^{\curlyvee})}{(\lambda_{BN}^{\curlyvee}-\lambda_{NN}^{\curlyvee})+(\lambda_{NP}^{\curlyvee}-\lambda_{BP}^{\curlyvee})}。$$

当 $\alpha_{\curlyvee}<\beta_{\curlyvee}$ 时，三支决策转换为二支决策，阈值 $\gamma_{\curlyvee}=\dfrac{(\lambda_{PN}^{\curlyvee}-\lambda_{NN}^{\curlyvee})}{(\lambda_{PN}^{\curlyvee}-\lambda_{NN}^{\curlyvee})+(\lambda_{NP}^{\curlyvee}-\lambda_{PP}^{\curlyvee})}$，则对于 $\forall X\in U$，其在不完备混合属性子集 B 下的“乐观”二支决策模型定义为

若 $\gamma_{\curlyvee}\le P(X|NR_B(x))\le 1$　则 $x\in POS_B^{NR}(X)$

若 $0\le P(X|NR_B(x))<\gamma_{\curlyvee}$　则 $x\in NEG_B^{NR}(X)$

根据定义 12，给出三支决策模型的语义解释：当 $x\in POS_B^{NR}(X)$ 时，代表接受 x 于 X 类中；当 $x\in BND_B^{NR}(X)$ 时，代表延迟接受或延迟拒绝 x 于 X 类中；当 $x\in NEG_B^{NR}(X)$ 时，代表拒绝 x 于 X 类中。

定义 13 在不完备混合决策系统 $IDS=U,C,D,\lambda_{\cdot\cdot}$ 中，令属性子集 $B=C$，$\forall X\in U/D$。则在“乐观”决策下，基于属性全集 C 对决策属性 D 的新的完备容差邻域关系下的正域、负域和边界域为

$$POS_C^{(\alpha_{\curlyvee},\gamma_{\curlyvee},\beta_{\curlyvee})}(D)_{NR}=\bigcup_{X\in U/D}POS_C^{(\alpha_{\curlyvee},\gamma_{\curlyvee},\beta_{\curlyvee})}(X)_S$$

$$BND_C^{(\alpha_{\curlyvee},\gamma_{\curlyvee},\beta_{\curlyvee})}(D)_{NR}=\bigcup_{X\in U/D}BND_C^{(\alpha_{\curlyvee},\gamma_{\curlyvee},\beta_{\curlyvee})}(X)_S$$

$$NEG_C^{(\alpha_{\curlyvee},\gamma_{\curlyvee},\beta_{\curlyvee})}(D)_{NR}=\bigcup_{X\in U/D}NEG_C^{(\alpha_{\curlyvee},\gamma_{\curlyvee},\beta_{\curlyvee})}(X)_S$$

由于“折中”和“悲观”决策风险下的三支决策和二支决策的基本模型与“乐观”决策风险下的基本模型类似，为了文章结构简洁，此处省略。

3　算法描述与实例分析

3.1　算法描述

在面向不完备混合决策系统的三支决策算法中，核心步骤为条件概率的计算和阈值的求解：首先，本文构建新的完备邻域容差关系获取不完备混合数据的近似类，并利用近似类计算条件概率；然后，为各对象设置不同损失函数区间，并通过新的区间值计算范式求解各对象的“乐观”“折中”和“悲观”三支阈值；最终构建基于新的完备邻域容差关系的“乐观”“折中”和“悲观”三支决策规则。算法描述如下：

算法：基于新的完备邻域容差关系三支决策规则

输入：不完备混合决策系统 IDS。

输出：“乐观”、“折中”和“悲观”三支决策规则。

Begin

对 IDS 中连续型数据进行标准化和归一化

初始化参数 ϑ、μ 和损失函数区间值 $\lambda_{\bullet\bullet}=[\lambda^{-},\lambda^{+}]$；

计算 $\delta_{a_k}=\dfrac{std_{a_k}}{\vartheta}$；//计算连续型属性下的 δ_{a_k} 值

计算决策类 $X\in U/D$；

for　$x\in U$　do

for　$y\in U$　do

基于新的完备邻域容差关系求解不完备混合数据的近似类 $NR_B(x)$；

end　//根据定义 9

end
根据 $NR_B^{\delta}(x)$ 计算 $\{\beta,\gamma,\alpha\}$；//根据定义 11
for　$x\in X$　do
计算的条件概率 $P(X|NR_B(x))$
end
for　$x\in X$　do
if　$(\beta\le\alpha)$
若 $1\ge P(X|NR_B(x))\ge\alpha$，则将 x 划分到类别 X 的正域 $POS_B^{NR}(X)$；
否则，若 $\beta<P(X|NR_B(x))<\alpha$，则将对象 x 划分到 X 的边界域 $BND_B^{NR}(X)$；
否则，将对象 x 划分到 X 的负域 $NEG_B^{NR}(X)$；
else
若 $\gamma\le P(X|NR_B(x))\le 1$　则将 x 划分到类别 X 的正域 $x\in POS_B^{NR}(X)$；
否则，则将 x 划分到类别 X 的负域 $x\in NEG_B^{NR}(X)$；
end
输出：$POS_C^{(\beta,\gamma,\alpha)}(D)_{NR}$、$POS_C^{(\beta,\gamma,\alpha)}(D)_{NR}$ 和 $POS_C^{(\beta,\gamma,\alpha)}(D)_{NR}$；//根据定义 13
end

其中 $\{\beta,\gamma,\alpha\}=\{\{\beta_{\curlyvee},\gamma_{\curlyvee},\alpha_{\curlyvee}\},\{\beta_{\prec},\gamma_{\prec},\alpha_{\prec}\},\{\beta_{\curlywedge},\gamma_{\curlywedge},\alpha_{\curlywedge}\}\}$；

算法时空复杂度分析：

算法的 Step1、Step3 的时间复杂度为 $O(|U||C_N|)$，C_N 代表连续型属性个数；Step5 的时间复杂度为 $O(|U|^2|C|)$；Step4、Step6、Step7、Step8 的时间复杂度为 $O(|U|)$，其余步骤的时间复杂度均为 $O(1)$。因此，该算法的最坏时间复杂度是 $O(|U|^2|C|)$；此外，存储空间主要用于存放不完备混合决策系统中的数据，故算法的空间复杂度为 $O(|U||C|)$。

3.2　实例分析

假定从某医院科室获取了一份包含不完备混合数据的诊断决策表 $IDS=U,C,V,D,\lambda_{.}$，如表 2 所示，其中 $U=\{x_1,x_2,x_3,x_4,x_5,x_6\}$ 代表 6 位患者；$C=\{a_1,a_2,a_3,a_4\}$ 代表四种检查项目：a_1—“咳嗽”、a_2—“过敏程度”、a_3—“胸闷程度”、a_4—“失眠程度”；V 代表检查结果，其中“1”—“有”、“0”—“没有”、“high”—“严重”、“low”—“不严重”，“*”表示因特殊原因暂时无法获取的数据，小数代表对应症状的患病程度；D 为诊断结果，符号“Φ”—“扁桃体炎症”、“Ψ”—“病毒感染”、“Ω”—“过敏”；$\lambda_{.}$ 代表各对象的损失函数区间，例如，对患者 x_1 而言，λ_{PP}、λ_{BP} 和 λ_{NP} 分别表示确诊其为扁桃体炎症患者后，对其进行治疗、延迟治疗和不治疗可能带来的风险损失；λ_{PN}、λ_{BN} 和 λ_{NN} 分别表示确诊其未患扁桃体炎症后，对其进行治疗、延迟治疗和不治疗可能带来的风险损失；其中“治疗”造成的损失主要包括检查费用和治疗费用等，“延迟治疗”造成的损失主要包括检查费用、治疗费用和病人因延迟治疗造成的身体损失等，“不治疗”造成的损失主要包括检查费用和病痛代价。通常情况下，$\lambda_{PP}\le\lambda_{BP}\le\lambda_{NP}$，$\lambda_{NN}\le\lambda_{BN}\le\lambda_{PN}$。

表 2 中“ω”为不同环境下的损失单位，比如此处可表示“千元”。根据决策属性划分决策类：$D_1=\{x_1,x_2,x_3\}$；$D_2=\{x_4\}$；$D_3=\{x_5,x_6\}$。令 $\vartheta=1$，计算得到表 2 中连续属性的邻域参数：$\delta_{a_2}\approx 0.23$，$\delta_{a_4}\approx 0.29$；表 2 的完备度：$\gamma\approx 0.79$。

根据定义 9，基于新的完备邻域容差关系对表 1 进行近似度量得到各对象的近似类：

新完备邻域容差类：$NR_C(x_1)=\{x_1\}$；$NR_C(x_2)=\{x_2,x_5\}$；$NR_C(x_3)=\{x_3\}$；$NR_C(x_4)=\{x_4\}$；$NR_C(x_5)=\{x_2,x_5\}$；$NR_C(x_6)=\{x_6\}$。

根据定义 11 计算患者的损失函数具体值，即分别设置 $\mu=0$、$\mu=0.5$ 和 $\mu=0.5$，再根据定义 12，计算“乐观”、“折中”和“悲观”三支阈值，如表 3~5 所示。

根据定义 12 计算出患者确诊患病的条件概率，并根据表 3~5 计算而来的三支阈值，对患者进行分类诊断，并给出治疗方案如表 6 所示，其中“？”表示不确定患者是否患有对应疾病。

表 2　某医院不完备混合数据诊断决策表

Table 2　Incomplete composite data diagnosis decision table of hospital

U	a_1	a_2	a_3	a_4	D	λ_{PP}	λ_{BP}	λ_{NP}	λ_{NN}	λ_{BN}	λ_{PN}
x_1	1	0.2	high	0.15	Φ	$[1\omega,2\omega]$	$[2\omega,3\omega]$	$[4\omega,5\omega]$	$[1\omega,3\omega]$	$[3\omega,4.5\omega]$	$[5\omega,7\omega]$
x_2	0	*	low	0.7	Φ	$[0\omega,2\omega]$	$[2.5\omega,3\omega]$	$[3\omega,5.5\omega]$	$[0.5\omega,1\omega]$	$[1.5\omega,3\omega]$	$[4\omega,5.5\omega]$
x_3	*	0.5	*	0.2	Φ	$[2\omega,3\omega]$	$[3\omega,3.5\omega]$	$[4\omega,5.5\omega]$	$[0\omega,1.5\omega]$	$[2\omega,3\omega]$	$[3.5\omega,5\omega]$
x_4	0	0.7	low	0.3	Ψ	$[0.5\omega,2\omega]$	$[4\omega,4.5\omega]$	$[5\omega,6\omega]$	$[0.5\omega,2\omega]$	$[2.5\omega,3\omega]$	$[3\omega,6\omega]$
x_5	0	0.8	low	0.8	Ω	$[1\omega,2.5\omega]$	$[3\omega,5\omega]$	$[5\omega,7\omega]$	$[1\omega,3\omega]$	$[3\omega,4\omega]$	$[4\omega,6\omega]$
x_6	0	*	*	0.85	Ω	$[1\omega,1.5\omega]$	$[2.5\omega,4\omega]$	$[4.5\omega,6\omega]$	$[2\omega,3\omega]$	$[4\omega,5.5\omega]$	$[6\omega,7.5\omega]$

表 3　各患者的“乐观”三支阈值的计算过程

Table 3　Calculation process of “optimistic” three-way thresholds of each patient

患者	“乐观”损失函数值						“乐观”三支阈值		
U	$\lambda_{PP}^{\curlyvee}$	$\lambda_{BP}^{\curlyvee}$	$\lambda_{NP}^{\curlyvee}$	$\lambda_{NN}^{\curlyvee}$	$\lambda_{BN}^{\curlyvee}$	$\lambda_{PN}^{\curlyvee}$	$\beta_{\curlyvee}$	$\gamma_{\curlyvee}$	$\alpha_{\curlyvee}$
x_1	1	2	4	1	3	5	0.50	0.57	0.67
x_2	0.5	2.75	4	0.75	2.25	4	0.55	0.48	0.44
x_3	2	3	4	0	2	3.5	0.67	0.64	0.60
x_4	0.5	4	5	0.5	2.5	3	0.67	0.36	0.13
x_5	0.5	2.75	4	0.75	2.25	4	0.55	0.48	0.44
x_6	1	2.5	4.5	2	4	6	0.50	0.53	0.57

表 4　各患者“折中”三支阈值的计算过程

Table 4　Calculation process of “compromising” three-way thresholds of each patient

患者	“折中”损失函数值						“折中”三支阈值		
U	$\lambda_{PP}^{\succ}$	$\lambda_{BP}^{\succ}$	$\lambda_{NP}^{\succ}$	$\lambda_{NN}^{\succ}$	$\lambda_{BN}^{\succ}$	$\lambda_{PN}^{\succ}$	$\beta_{\succ}$	$\gamma_{\succ}$	$\alpha_{\succ}$
x_1	1.5	2.5	4.5	2	3.75	6	0.47	0.57	0.69
x_2	1.38	3.38	5.13	1.38	2.88	4.88	0.46	0.48	0.50
x_3	2.5	3.25	4.75	0.75	2.5	4.25	0.54	0.61	0.70
x_4	1.25	4.25	5.5	1.25	2.75	4.5	0.55	0.43	0.37
x_5	1.38	3.38	5.13	1.38	2.88	4.88	0.46	0.48	0.50
x_6	1.25	3.25	5.25	2.5	4.75	6.75	0.53	0.52	0.50

表 5　各患者"悲观"三支阈值的计算过程

Table 5　Calculation process of "pessimistic" three-way thresholds of each patient

患者	"悲观"损失函数值						"悲观"三支阈值		
U	$\lambda_{PP}^{\wedge}$	$\lambda_{BP}^{\wedge}$	$\lambda_{NP}^{\wedge}$	$\lambda_{NN}^{\wedge}$	$\lambda_{BN}^{\wedge}$	$\lambda_{PN}^{\wedge}$	$\beta_{\wedge}$	$\gamma_{\wedge}$	$\alpha_{\wedge}$
x_1	2	3	5	3	4.5	7	$\underline{0.43}$	0.57	$\underline{0.71}$
x_2	2.25	4	6.25	2	3.5	5.75	$\underline{0.40}$	0.48	$\underline{0.56}$
x_3	3	3.5	5.5	1.5	3	5	$\underline{0.43}$	0.61	$\underline{0.80}$
x_4	2	4.5	6	2	3	6	$\underline{0.40}$	0.48	$\underline{0.55}$
x_5	2.25	4	6.25	2	3.5	5.75	$\underline{0.40}$	0.48	$\underline{0.56}$
x_6	1.5	4	6	3	5.5	7.5	0.56	$\underline{0.52}$	0.44

在本例中，由于论域 U 较小，所以各对象的在新的完备邻域容差关系下的近似类也较小，大多甚至只有本身，这使得条件概率大多都为 1，进而使得大部分对象不论在"乐观"诊断下还是"折中"诊断下都被划分到正域并被推荐治疗；同时可以看到，在"悲观"诊断下，患者 x_2 和 x_5 被划分到了边界域中，即不确定 x_2 是否患有扁桃体炎症，不确定 x_5 是否患有过敏，都需要进一步诊断，因此，三支决策为真实决策过程提供了修正错误分类的方法，基于新的完备邻域容差关系适用于医疗诊断这类不完备混合决策系统，而且相比于下文所提的其他方法更加客观合理。

表 6　基于三支决策为各患者制定三种决策风险下的治疗方案

Table 6　Therapeutic schedule is provided based on three-way decisions for each patient

患者	条件概率	"乐观"医疗诊断			"折中"医疗诊断			"悲观"医疗诊断		
U	$P(D\|NR_B(x))$	划分结果	诊断结果	治疗方案	划分结果	诊断结果	治疗方案	划分结果	诊断结果	治疗方案
x_1	1	$POS_C^{NR}(D_1)$	扁桃体炎症	治疗	$POS_C^{NR}(D_1)$	扁桃体炎症	治疗	$POS_C^{NR}(D_1)$	扁桃体炎症	治疗
x_2	0.5	$POS_C^{NR}(D_1)$	扁桃体炎症	治疗	$POS_C^{NR}(D_1)$	扁桃体炎症	治疗	$BND_C^{NR}(D_1)$	扁桃体炎症？	延迟治疗
x_3	1	$POS_C^{NR}(D_1)$	扁桃体炎症	治疗	$POS_C^{NR}(D_1)$	扁桃体炎症	治疗	$POS_C^{NR}(D_1)$	扁桃体炎症	治疗
x_4	1	$POS_C^{NR}(D_2)$	病毒感染	治疗	$POS_C^{NR}(D_2)$	病毒感染	治疗	$POS_C^{NR}(D_2)$	病毒感染	治疗
x_5	0.5	$POS_C^{NR}(D_3)$	过敏	治疗	$POS_C^{NR}(D_3)$	过敏？	治疗	$BND_C^{NR}(D_3)$	过敏？	延迟治疗
x_6	1	$POS_C^{NR}(D_3)$	过敏	治疗	$POS_C^{NR}(D_3)$	过敏	治疗	$POS_C^{NR}(D_3)$	过敏	治疗

为了与新的完备邻域容差关系进行比较，利用邻域容差关系 NT、限制邻域容差关系 NL、完备邻域容差关系 NM 对表 2 中的患者进行近似划分和分类诊断。详细计算结果如下：

基于邻域容差关系

根据定义 2 计算邻域容差类：

$NT_C(x_1)=\{x_1\}$；$NT_C(x_2)=\{x_2,x_5,x_6\}$；$NT_C(x_3)=\{x_3,x_4\}$；$NT_C(x_4)=\{x_3,x_4\}$；$NT_C(x_5)=\{x_2,x_5,x_6\}$；$NT_C(x_6)=\{x_2,x_5,x_6\}$；

根据定义 11 计算基于邻域容差类的三支损失函数值，然后根据定义 3 划分三种决策风险下的正域、边界域和负域：

"乐观"：$POS_C^{NT}(D)=\{x_1,x_3,x_4,x_5,x_6\}$；$BND_C^{NT}(D)=\{\varnothing\}$；$NEG_C^{NT}(D)=\{x_2\}$；

"折中"：$POS_C^{NT}(D)=\{x_1,x_4,x_5,x_6\}$；$BND_C^{NT}(D)=\{\varnothing\}$；$NEG_C^{NT}(D)=\{x_2,x_3\}$；

"悲观"：$POS_C^{NT}(D)=\{x_1,x_5,x_6\}$；$BND_C^{NT}(D)=\{x_3,x_4\}$；$NEG_C^{NT}(D)=\{x_2\}$；

分析结果发现：尽管患者 x_6 的"过敏程度"和"胸闷程度"都未检测到，但在邻域容差关系下却被划分到了 x_2 的"扁桃体患者"一类中，导致 x_2 被划分到负域中，间接就影响到了 x_2 的治疗方案—不治疗，而实际上，x_2 的检测项目中有 75%是已知的，应该被推荐治疗或延迟治疗。同样的情况也发生在 x_5 上，因此，对于不对未知值做任何处理的邻域容差关系而言，并不适用于类似于医疗诊断中伴有缺失数据的混合决策系统。相反，在新的完备邻域容差关系下进行分类诊断就不会出现此类不合理现象。

经过详细计算，在本例中由于实例对象较少，在限制邻域容差关系 NL 和完备邻域容差关系 NM 下的近似度量结果与在邻域容差关系 NT 的近似度量结果一致，所以三域划分结果也相同。为避免重复，此处省略。同时也证明了 NL 和 NM 也不适用于类似于医疗诊断中出现的不完备混合决策系统。

通过该医疗诊断的实例证明，基于新的完备邻域容差关系的三支决策模型能客观合理地对患者进行医疗诊断，证明了新的完备邻域容差关系对于不完备混合数据处理的合理性和可解释性，扩充了三支决策对此类信息系统的研究范围。

4　实验分析与对比

为进一步验证本文方法对处理不完备混合决策系统的有效性和可行性，选取六个 UCI 混合型数据集进行实验分析。数据集的描述如表 7 所示。实验运行环境为：Win10，Intel(R) Core(TM)，CPU i5-6300HQ 2.30 GHz 和 8.0 GB 内存，用 Python 编程语言在开发平台 Pycharm 2017.3.4 (community edition)上实现。为了便于实验分析和比较，用 NTTWD 代表邻域容差关系下的三支决策模型，NLTWD 代表限制邻域容差关系下的三支决策模型，NMTWD 代表完备邻域容差关系下的三支决策模型。本文方法—NRTWD 代表新的完备邻域容差关系下的三支决策模型；为保证实验的公平性，以上方法均在曼哈顿距离下进行计算，同时为了消除量纲的影响，对所有的连续型数据进行标准化和归一化，并对同一份数据只通过"折中"阈值设置法做一次阈值的获取。实验将从以下度量函数指标进行分析和比较：

准确率为 $\mathrm{Acc}=\dfrac{n_{pp}}{n_{pp}+n_{np}}$。

覆盖率为 $\mathrm{Cov}=\dfrac{n_{pp}+n_{np}}{n_{pp}+n_{bp}+n_{np}}$。

权衡因子为 $\mathrm{F}=2*\dfrac{Acc*Cov}{Acc+Cov}$。

误划分损失为 $\mathrm{Cost}=n_{bp}*\lambda_{bp}+n_{np}*\lambda_{np}$。

其中：F 代表分类能力；设 n_{pp}、n_{bp}、n_{np} 分别为正域、边界域和负域中的对象数；λ_{bp} 和 λ_{np} 分别为当对象实际属于某类别时被划分到该类别的边界域和负域所造成的损失。以下实验本文均设置 $\lambda_{bp}=0.3$，$\lambda_{np}=0.7$。

表 7　UCI 数据集描述

Table 7　Descriptions of UCI datasets

数据集	对象数	属性数			类别数	缺失程度
		字符型	整型	连续型		
Heart	270	0	8	5	2	0%
Automobile	205	10	1	15	6	1.2%
Credit Approval	690	9	0	6	2	0.65%
Cylinder bands	540	20	0	20	2	4.73%
Hepatitis	155	0	13	6	2	5.67%
Horse Colic	368	0	17	10	2	19.39%

4.1　参数 ϑ 对本文方法的单调性影响

由定义 9 可知，ϑ 的取值决定了邻域参数 δ 的大小，从

而影响近似类的大小，最终影响 NRTWD 的分类结果，因此，本节将从表 7 中选取后五个真实缺失的混合数据集进行实验分析，讨论当 ϑ 从 0.1 到 1.0 逐渐递增时，NRTWD 的准确率和误分类损失的变化趋势。

ϑ 对 NRTWD 的划分准确率的影响如图 1 所示。由图 1 可知，NRTWD 的划分准确率大致随着 ϑ 的逐渐增大而增大，除 Horse Colic 数据集外，其余数据集的 Acc 整体变化跨度不大，当 ϑ 从 0.5 增大到 0.6 时，Automobile 数据集下的 Acc 从 0.946 3 增大到 0.961 0，Credit Approval 从 0.960 9 增大到 0.972 5，Horse Colic 从 0.913 0 增大到 0.934 8。但仔细观察发现，当 ϑ 从 0.1 增大到 0.2 时，Cylinder bands 下的 Acc 却从 0.979 6 降到了 0.944 4，类似地，该数据在 $\vartheta=0.5\sim0.6$ 间也略微下降了，可见 ϑ 的变化不会引起 NRTWD 的分类精度的严格地单调性变化。ϑ 对 NRTWD 的误分类损失的影响如图 2 所示。由图 2 可知，在大多数数据集下，NRTWD 的误分类损失随着 ϑ 逐渐增大而大致呈现下降趋势。例如，ϑ 从 0.1 增大到 1.0 的整个过程中，Horse Colic 下的 Cost 从 42.7 不断降到了 11.2，Credit Approval 从 25.9 降到了 8.4，Automobile 从 9.8 降到了 2.1，Hepatitis 从 7.7 降到了 0.7，而 Cylinder bands 下的 Cost 却在 $\vartheta=0.1\sim0.2$ 时突然从 7.7 上升到了 21，类似情况也发生在 $\vartheta=0.5\sim0.6$ 间，由此可见，ϑ 的变化也不会对 NRTWD 的误分类损失造成严格的单调变化。但总的来说，ϑ 越大，NRTWD 的 Acc 越高，Cost 越低。

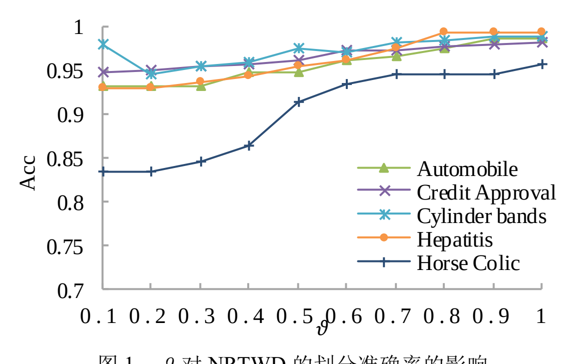


图 1　ϑ 对 NRTWD 的划分准确率的影响

Fig.1　Acc curves of NRTWD varying with ϑ

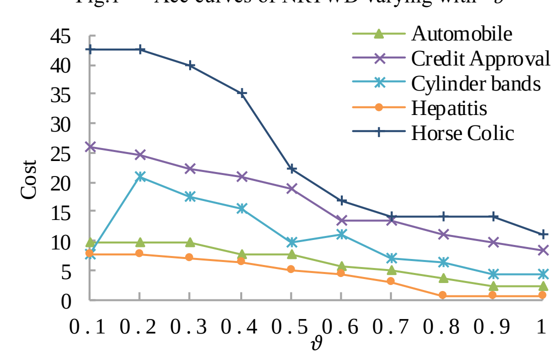


图 2　ϑ 对 NRTWD 的误分类损失的影响

Fig.2　Cost curves of NRTWD varying with ϑ

4.2　分类性能分析与比较

本节主要对比五种方法的分类性能。首先，对无缺失的 Heart 数据集进行随机缺失 5%~30%处理，令 $\vartheta=1$，通过 NTTWD、NLTWD、NMTWD 和 NRTWD 方法对其进行实验分析，并将 Acc、F 和 Cost 值展示在表 8 中。然后，通过以上四种方法对表 6 中除 Heart 数据集以外的其他真实缺失的数据集进行实验，并将各方法在 $\vartheta=0.6\sim1$ 下的 Acc、F 和 Cost 的平均值展示在表 9 中。

如表 8 所示，对于随机缺失的 Heart 数据集，从整体情况分析，随着缺失程度不断提高，各方法的 Acc 和 F 都不断下降，Cost 不断提高，由此可见，随着缺失程度逐渐提高，各方法的分类性能在逐渐降低；但仔细分析发现，NRTWD 的 Acc 和 F 值始终要高于、Cost 值始终要低于 NTTWD、NLTWD 和 NMTWD，由此可知，NRTWD 受缺失程度的影响要小于其他三种方法。例如，当 Heart 从缺失 5%到 30%时，NTTWD、NLTWD 的 Acc 下降了 13.64%，F 下降了 7.41%，Cost 值上升 25.9，NMTWD 的 Acc 和 F 分别下降了 15.56%和 8.5%，而 NRTWD 的 Acc 只下降 10%，F 只下降 5.3%，Cost 只上升 18.9。另外可以看到，NTTWD 和 NLTWD 对于 heart 数据集的分类效果一致，一定程度上说明这两种方法对于不完备混合数据集的分类效果相差不大。

从表 9 看出，对于真实缺失的 UCI 数据集，在缺失程度低的 Automobile 和 Credit Approval 数据集下，NRTWD 的 Acc、F 和 Cost 要优于 NTTWD、NLTWD 和 NMTWD；而在缺失程度高的 Cylinder bands、Hepatitis 和 Horse Colic 数据集下，NRTWD 略低于 NTTWD、NLTWD 和 NMTWD，但仍然保持较高的分类效果。同时也可以看到，NTTWD 和 NLTWD 的划分结果在以上所有数据集上是相同的。

综上所述，从理论上分析 NLTWD 要略优于 NTTWD，但实验结果显示 NTTWD 和 NLTWD 方法对于大多数缺失程度或大或小的不完备混合数据集的分类结果基本都一致；NRTWD 对于同一数据集的分类性能会随该数据的缺失程度逐渐增大而逐渐降低，但始终优于 NTTWD、NLTWD 和 NMTWD，而对于真实缺失的数据集，NRTWD 的分类性能在大多数情况下要优于其他方法，因此，结合理论分析和实验对比可知，对于不完备混合数据的处理，NRTWD 的分类过程比其他方法更合理，分类效果更优。

表 8　各方法在 Heart 数据集下的分类性能对比

Table 8　Comparison of classification performance of each method under heart dataset

度量	不同关系	缺失程度				
		5%	10%	15%	25%	30%
Acc	NTTWD	0.9926	0.9481	0.9519	0.8630	0.8556
	NLTWD	0.9926	0.9481	0.9519	0.8630	0.8556
	NMTWD	0.9926	0.9444	0.9185	0.8407	0.8370
	NRTWD	0.9926	0.9593	0.9593	0.8741	0.8926
F	NTTWD	0.9963	0.9734	0.9754	0.9265	0.9222
	NLTWD	0.9963	0.9734	0.9754	0.9265	0.9222
	NMTWD	0.9963	0.9714	0.9575	0.9135	0.9113
	NRTWD	0.9963	0.9792	0.9792	0.9328	0.9433
Cost	NTTWD	1.4	9.8	9.1	25.9	27.3
	NLTWD	1.4	9.8	9.1	25.9	27.3
	NMTWD	1.4	10.5	15.4	30.1	30.8
	NRTWD	1.4	7.7	7.7	23.8	20.3

5　结束语

三支决策作为知识发现和人工智能领域中一种粒计算方法，近年来许多研究人员对三支决策理论及其应用开展了深入研究。本文分析了不同邻域容差关系下的三支决策模型，并在此基础上，提出了一种新的完备邻域容差关系为不完备混合决策系统构建三支决策模型，定义新的损失函数值计算公式用于获取各对象的乐观、折中和悲观三支阈值，最后给出了面向不完备混合决策系统的乐观、折中和悲观三支决策规则。通过医疗诊断的实例和实验证明，对于不完备混合决策系统的处理，本文所提的三支决策模型较其他模型更加客

观合理，分类效果更优；同时，该模型也一定程度上丰富了三支决策的理论研究，合理解释了三支决策在应用领域中的实践意义。在下一步的工作中，本文将研究不完备混合数据系统下的三支属性约简模型，并探索其应用范围。

表 9　各方法在真实缺失的数据集下的分类性能对比

Table 9　Comparison of classification performance of each method under each truly missing dataset

度量	数据集	NTTWD	NLTWD	NMTWD	**NRTWD**
Acc	Automobile	0.9571	0.9571	0.9571	0.9747
	Credit Approval	0.9739	0.9739	0.9742	0.9803
	Cylinder bands	1.0000	1.0000	0.9837	0.9837
	Hepatitis	0.9974	0.9974	0.9780	0.9832
	Horse Colic	1.0000	1.0000	0.8745	0.9457
F	Automobile	0.9780	0.978	0.9780	0.9871
	Credit Approval	0.9868	0.9868	0.9869	0.9900
	Cylinder bands	1.0000	1.0000	0.9918	0.9918
	Hepatitis	0.9987	0.9987	0.9888	0.9915
	Horse Colic	1.0000	1.0000	0.9329	0.9721
Cost	Automobile	6.16	6.16	6.16	3.64
	Credit Approval	12.6	12.6	12.46	9.52
	Cylinder bands	0	0	6.16	6.16
	Hepatitis	0.28	0.28	2.38	1.82
	Horse Colic	0	0	32.34	14

参考文献：

[1] Yao Y Y. The superiority of three-way decisions in probabilistic rough set models [J]. Information Sciences, 2011, 181 (6): 1080-1096.

[2] Fujita H, Li Tianrui, Yao Y Y. Advances in three-way decisions and granular computing [J]. Knowledge-Based Systems, 2016, 91: 1-3.

[3] Wang Pingxin, Yao Y Y. CE3: A three-way clustering method based on mathematical morphology [J]. Knowledge-Based Systems, 2018, 155 (1) : 54-65.

[4] Hu Baoqing, Wong H, Yiu K F C. On two novel types of three-way decisions in three-way decision spaces [J] , International Journal of Approximate Reasoning, 2017, 82: 285-306.

[5] Chen Yumin, Zeng Zhiqiang, Zhu Qing-xin, *et al.* Three-way decision reduction in neighborhood systems [J]. Applied Soft Computing, 2016, 38 (1): 942-954.

[6] Zhang Xianyong, Miao Duoqian. Three-way attribute reducts [J]. International Journal of Approximate Reasoning, 2017, 88: 401-434.

[7] Yang Xin, Li Tianrui, Fujita H, *et al.* A unified model of sequential three-way decisions and multilevel incremental processing [J]. Knowledge-Based Systems, 2017, 134: 172-188.

[8] 刘久兵，张里博，周献中，等. 直觉模糊信息系统下的三支决策模型 [J]. 小型微型计算机系统，2018，39 (6)：1281-1285. (Liu Jiubing, Zhang Libo, Zhou Xianzhong , *et al.* Three-way decision model under intuitionistic fuzzy information system environment [J]. Journal of Chinese Computer Systems, 2018, 39 (6): 1281-1285.)

[9] Yao Y Y, Wang Shu, Deng Xiaofei. Constructing shadowed sets and three-way approximations of fuzzy sets [J]. Information Sciences, 2017, (412-413): 132-153.

[10] Liang Deicui, Liu Dun. Systematic studies on three-way decisions with interval-valued decision-theoretic rough sets [J]. Information Sciences, 2014, 276 (8): 186-203.

[11] Zhang Hengru, Min Fan. Three-way recommender systems based on random forests [J] , Knowledge-Based Systems, 2016, 91 (1) : 275-286.

[12] Li Huaxiong, Zhang Libo, Huang Bing, *et al.* Sequential three-way decision and granulation for cost-sensitive face recognition [J] , Knowledge-Based Systems, 2016, 91 (1): 241-251.

[13] Nauman M, Azam N, Yao Jingtao. A three-way decision making approach to malware analysis using probabilistic rough sets [J]. Information Sciences, 2016, 374: 193-209, .

[14] Fernandes V, Yevseyeva I. Mendez JR, *et al.* A spam filtering multi-objective optimization study covering parsimony maximization and three-way classification [J]. Applied Soft Computing, 2016, 48: 111-123.

[15] Kryszkiewicz M, Rough set approach to incomplete information systems [J]. Information Sciences, 1998, 112 (1): 39-49.

[16] Stefanowski J, Tsoukiàs A. On the extension of rough sets under incomplete information [J]. International Journal of Intelligent System, 2000, 16 (1): 29-38.

[17] 王国胤. Rough 集理论在不完备信息系统中的扩充 [J]. 计算机研究与发展，2002, 39 (10): 1238-1243. (Wang Guoyin, Extension of rough set under incomplete information systems [J] . Journal of Computer Research and Development, 2002, 39 (10): 1238-1243.)

[18] 盛立，杨慧中. 基于完备容差关系的扩充粗糙集模型 [J]. 控制与决策，2008, 23 (3): 258-262. (Sheng Li, Yang Huizhong. Extended rough set model based on completed tolerance relation [J] . Control and Decision, 2008, 23 (3): 258-262.)

[19] 马希骜，王国胤，张清华，等. 基于改进的完备容差关系的扩充粗糙集模型 [J]. 计算机应用，2010, 30 (7): 1873-1877. (Ma Xiao, Wang Guoyin, Zhang Qinghua *et al.* Extended rough set model based on improved complete tolerance relation [J] . Journal of Computer Applications, 2010, 30 (7): 1873-1877.)

[20] Liu Dun, Liang Deicui, Wang Changchun. A novel three-way decision model based on incomplete information system [J] . Knowledge-Based Systems. 2016, 91 (1): 32-45.

[21] 姚晟，汪杰，徐风，等. 不完备邻域粗糙集的不确定性度量和属性约简 [J]. 计算机应用，2018, 38 (1): 97-10. (Yao Sheng, Wang Jie, Xu Feng, *et al.* Uncertainty measurement and attribute reduction in incomplete neighborhood rough set [J]. Journal of Computer Applications, 2018, 38 (1): 97-103.)

[22] Zhao Hua, Qin Keyun. Mixed feature selection in incomplete decision table [J]. Knowledge-Based Systems, 2014, 57 (2): 181-190.